Is the Human a Machine, or Is It Also Mind?

François Fleuret

June 2015

The development of computers and the constant increase in their complexity and computing power for more than six decades have reduced the gap between their "cognitive abilities" and those of humans. This evolution makes the difference between humans and machines increasingly blurred. In particular, it renders obsolete many arguments that relied on the obvious triviality of the "behavior" of machines.

1 Introduction

What is the reality of the substrate of the human mind? Is it a very complex clock, or does it involve "something more"? This is a nuanced and difficult question, to which yet the vast majority of people respond with certainty in the negative: No, the human is not just a "machine". As the title of this essay puts it, it is also "mind".

Just as it has been difficult to accept that man is the descendant of animals, or that life emerges from inert matter, it is shocking today to imagine that the psyche, feelings and self-awareness could emerge from electrical or mechanical operations. This opinion probably comes from a false preconception of what machines are, and implicitly of their limitations.

2 The Nature of the Question

The difficulty of the question comes in part from the complexity of the three terms it combines. If "human", "machine" and "mind" have roughly shared meanings in everyday discussions, they are not sufficient to analyze the question we are asking ourselves.

2.1 Machine

In the common sense, a machine is a device built by man that can use a source of energy to perform a task without human intervention: a windmill, a washing machine, or a clock.

The meaning of "device" is not clear, and the constraint that it must be "built by man" is excessive. We perfectly accept that machines are manufactured by other machines. If it is design and not construction that we are talking about, it is not clear either. Many machines - airplanes, engines, microprocessors - are now designed in part by computers that calculate optimal shapes and configurations.

For the discussion at hand, however, we must impose constraints on the components we accept in a machine. It is likely that we would not call "machine" the set composed of a mill and the animal that turns it. We cannot accept that complex structures are taken "from nature" and used as is. Such a hybrid could in particular include fragments of human brains and the question would lose its meaning.

2.2 Human

We are obviously not interested in the appearance or physical constitution of the human being, but in its "broad sense" cognitive abilities. That is to say, its ability to solve problems in the real world - whether intellectual or motor - but also to have a conscious experience.

The difficulty comes from this last point. We each have a subjective impression of being. But we cannot qualify the humanity of others according to their subjective experience. We cannot, neither in practice nor in principle, be in the place of others. We can imagine what they feel by using our own experience as a model, but we have no way of knowing if it is their experience.

The argument is essentially that even if you had access to the internal functioning of another person's brain, you would need to have exactly the same means as them to interpret said internal functioning, so this interpretation should be done by their brain. As long as a part of the processing is done by yours, the resulting interpretation has no reason to be correct. But if no fraction of your brain is involved, you do not feel things, the other feels them [8, 3].

The only alternative to define what a human is is therefore an external characterization. Not to define it according to its subjective experience, to which nothing and no one else will ever have access, but according to its behavior in the world. Such a characterization takes the form of a behavioral test such as the Turing test [12]. The principle of the latter is based on a test during which a human examiner communicates with an interlocutor that he neither sees nor hears. The interaction is therefore done for example using a keyboard and a screen. This examiner must guess at the end of a discussion with his interlocutor if the latter is a human or a machine that "pretends" to be a human. If the machine manages to deceive examiners more than half the time during such tests, it is considered to pass the test.

Deciding whether this test is sufficiently constrained - how many times must the machine pass the test? what is the intelligence and expertise of the examiners? - and whether it is sufficient is a philosophical problem that we will not solve here [4, 1].

Nevertheless, one should not underestimate the richness of such a test. The discussion with the machine can address questions related to emotions, self-awareness, and the introspective processes associated with it. To pass it, the machine must therefore at least know how to imitate introspective capacities and the ability to report a coherent subjective experience. In fact, we consider as acquired the humanity of the people with whom we interact every day, even though we only know them through a limited version of the Turing test.

2.3 Mind

The term "mind" is probably the most poorly defined of the three that interest us. It can refer to cognitive functions as well as to incorporeal or divine components of the human. The implicit meaning in the present discussion is the non-physical part of the human, therefore "what is not a machine".

This interpretation is dangerous because it implies that a machine exists only in physical reality. However, an important part of what makes modern machines useful, especially computer programs, is pure information, independent of the substrate in which it is represented. It can be duplicated, transmitted and encoded in various forms. The same calculation can be performed using an electronic, mechanical or hydraulic device, or by a human with a pencil and sheets of paper.

A definition of the mind - which would not be in opposition to "machine" - could therefore be that it is the "informational" model of a cognitive system, detached from its physical realization, like a musical score or a recipe are representations that can materialize into tangible objects. It is therefore immaterial and permanent.

2.4 What is the Question

These definitions being posed, the question therefore becomes: "Can there exist a computer that would pass the Turing test?" or "Can a human do something that a computer cannot do in principle?"

The conviction that the human can do more than a machine probably comes from a false idea of what a machine is, itself due to the observation of everyday machines. The latter represent only a tiny part of the set of machines that can exist in principle, and induce an erroneous intuition about them.

3 The Reality of Machines

Machines are perceived as being of limited complexity and incapable of going beyond the behaviors specified by their designers. Modern computers force us to reconsider this view.

3.1 Complexity

The term "machine" covers objects as diverse as a windmill, a loom, or a computer. However, the complexities of these different devices are of totally different orders.

A standard computer, as can be purchased in a supermarket, is equipped with a memory that contains several tens of billions of digits, and a computing unit capable of performing several tens of billions of operations per second. These values are not hyperbolic quantities with obscure meaning, but concrete specifications: A computer that costs a few hundred francs does multiplications a hundred times faster than humanity as a whole.

This immense complexity makes the nature of the process that takes shape in the calculations performed by a computer confusing.

A large number of simulation techniques, for example for weather forecasting, optimization of the resistance of car passenger compartments, or the design of aircraft fuselages, are based on the same idea of "finite elements", which consists in decomposing the object of interest into very small parts. The motivation behind this approach is that we know how a small volume of atmosphere or metal behaves when subjected to constraints, but we do not know what the overall behavior that results from it is. A computer, thanks to its computing power, can perform the same "simple" calculation billions of times, "small part by small part", and finally simulate an overall behavior. This makes it possible to answer questions about this overall behavior: "does the passenger compartment crush too much and put the passengers in danger?" or "with this wing shape, does the plane consume more fuel?"

When such a simulation runs in a computer, it results in two different levels of reality: that where the calculation itself is done, which is a succession of a very large number of simple calculations, and that where the object of interest exists. If we simulate a mechanical clock in a computer by treating small bits of matter, and "the hands turn", which concretely means that the numbers in the computer's memory fluctuate in a way that can be interpreted as such, what exactly are we talking about? Do these hands "exist"?

If we simulated a human brain in a computer, molecule by molecule, how would our understanding of the simulation mechanisms give us a valid intuition about the result of this simulation?

3.2 Predictability and Optimization

The second supposed limitation of machines is their predictability.

A trivial argument against this idea is that one can make non-deterministic machines. In the common sense by using chaotic mechanical processes as is done for a lottery machine, or in a more exact sense by using measurements of quantum phenomena.

This trick is not convincing, but it allows us to understand that it is not determinism in the strict sense that we reproach a machine for, but rather an "existence perimeter", from which we think it will never be able to escape. We

know what the past and future of a clock are, we know what trajectory it will follow forever.

Such a perimeter is much more difficult to define for a computer program because the latter does not follow a series of operations that are executed one after the other in a systematic way. It decides which operations to perform based on those it has previously executed. This "conditional" behavior is completely specified by the programmer, but it results in a complexity of operation much higher than the complexity of the program itself.

Combined with the immense computing power of computers, this behavior allows a computer program to produce in practice a result that goes beyond what its designers can anticipate. A chess program can find a move that no human could have imagined, and a robot arm control system can find a way to position a part in a factory assembly more efficiently than the best strategy imagined by humans before it.

More generally, most optimization programs - of which a chess program or a robot arm controller are two examples - are only of interest precisely because what they produce could not have been found by a human. They provide "better" solutions simply because they have at their disposal a computing power and memory capacity superior to their designers. By principle, their usefulness is directly linked to their unpredictability.

3.3 Statistical Learning

In all the examples that precede, a clear distinction persists between the programs, designed by humans, and the results that these programs produce. However, this distinction is arbitrary.

It very frequently happens that the functioning of a program depends on parameters which are empirically adapted according to the data it must process. This adaptation is very close to the mechanical adjustment of a machine to adapt it to the context of its use.

An important field of contemporary computer science, statistical learning, studies and develops methods that allow the adaptation of very large numbers of parameters. This type of approach is particularly useful for making predictions for which it is difficult to formally define a rule. For example, to automatically predict which object is visible in an image, to determine if a cell is cancerous from a signature of gene expressions, or to recognize a word in a sound recording. These tasks, although very easy for humans and animals when it comes to vision and hearing, have until recent years been the Achilles' heel of computers.

The principle of the most used method for these tasks consists in progressively modifying the parameters so that the program predicts what it must predict on "learning examples", for example images for which it has been indicated which objects are visible. For each of these examples, the computer makes a small correction to each parameter so that the predicted response is "closer" to the desired response [9].

The complexity of computers, combined with this type of learning method, makes the distinction between a program and the product of a program very

blurry. The number of parameters that such techniques can estimate today reaches several billion, and allows them to have universality properties. In practice, as long as it is provided with enough learning examples, such a program will learn how to make a correct prediction, whatever the complexity of the corresponding rule.

For some real applications such as language analysis or medicinal molecule selection, teams without domain expertise have obtained better results using these learning techniques than domain experts by explicitly constructing models [2, 5]. We thus arrive in such a case at a situation where a form of "understanding of the world" comes from a calculation and learning data, and in no case from the human designers of the program.

Moreover, if the structures obtained by learning do well what they are supposed to do, it is very difficult to understand how, and their analysis constitutes a research subject in its own right [13, 11, 15].

3.4 Emotions and Lies

What precedes shows that machines do not suffer from obvious limitations of principle that would prevent them from developing the same cognitive capacities as humans.

Nevertheless, one can wonder if, even if they could potentially develop the same capacities as a human, they would do so. This question is different, because it involves the context in which the machine evolves. Its objectives, its past, and its "culture". Two traits often come back as being inaccessible to machines: Emotions and lies.

If these traits are only important in the discussion because they reflect strong states of consciousness, and are used as concrete examples of what is meant by "self-awareness", then we come back to the problem addressed in § 2.2. We will never be able to know if a machine, or another human than ourselves, who expresses an emotion or who lies, actually feels something particular.

If it is the behavioral aspect that is at issue, there is no impossibility of principle. One can perfectly imagine that a machine, for example an "intelligent robot", could have fundamental states corresponding to fear, or joy, when it is in a situation of danger for its physical integrity, or on the contrary when it reaches one of the fundamental goals for which it hasbeen trained [6]. As for lies, they constitute the best strategy in the case where the machine exchanges information and is in competition with third parties for a limited resource. Such behaviors appear naturally in simulations with robots [7].

4 Conclusion

The objective of this essay is essentially to refute "common sense" arguments that are false. Thanks to the extreme complexity of modern computers, the machines they equip differ qualitatively from those of everyday life. In particular, they can produce unexpected results, and autonomously modify their

functioning according to their interactions with the environment.

Man is probably a machine, but the possible and future machines are much more than what our intuition tells us.

A Responses to Wulfram Gerstner

A.1 Machines Do Not Have Introspective Capacities

There is no impossibility of principle that a machine is introspective, that is to say that it has access to its internal representations as it has access to observations on the external world.

A.2 A Simulation Is Not Reality

There is no clear separation between simulations and reality. There exist for example hearing aids or artificial retinas that replace nervous tissue. An information processing that normally takes place in neurons is thus carried out using calculations in microprocessors, and leads to the same conscious experience for the one who wears the prosthesis. Is this treatment a simulation? Or a physical reality?

Wulfram cites the example of computer simulations in which simulated water molecules take configurations similar to ice, and he points out that this simulated ice is not cold. This argument is circular in the debate at hand: A conscious artificial intelligence that would interact with the "simulated" ice would feel cold, so this ice would be cold.

A.3 The Chinese Room Demonstrates the Insufficiency of the Turing Test

The Chinese room is a central "thought experiment" in debates about consciousness, and serves to demonstrate the insufficiency of the Turing test [10]. In this imaginary experiment one or more individuals are in a closed room, and interact with an outside interlocutor in a language they do not understand. They refer to documents that formally indicate, for each sentence that is said to them, what they should answer.

There are two ways to understand this thought experiment, and both are in my opinion unsatisfactory.

The first is to say that since none of the individuals involved is aware of the content of the discussion, but that this discussion allows to pass the Turing test, then one can indeed pass this test without a conscious state. This first interpretation confuses the state of consciousness of the elementary components of a system and the state of consciousness of the entire system. Similarly, one would demonstrate that the brain is not conscious because neurons cannot be individually.

The second interpretation of this experiment is more generally that a very simple system, and therefore obviously not conscious, can pass the Turing test. One could replace the individuals with a primitive computer program that "looks in a book" for the sentence to answer.

But for combinatorial reasons this book cannot exist. The size of the universe would not allow to make an object that would contain all the discussions necessary to pass the Turing test. The counter-argument is then that the book can be replaced by a more complex device to "compress the information", for example by grouping identical sentences, or by using synonym tables. But in this case, the more the system becomes achievable, the more complex it is, and the less the simplicity argument holds.

A.4 The Moral Implications Are Unacceptable

Wulfram's last argument, put forward in his rejoinder, is based on the moral consequences of this debate. If humans are machines, what is the gravity of a murder?

The first counter-argument is that the universe is not moral, and that moral consequences reveal nothing about the physical reality of the world. The second counter-argument is that deciding what is the gravity of the destruction of a machine is precisely the question at hand: If humans are machines, then there exist machines whose destruction is an extremely serious moral act.

B Responses to the Public

B.1 A Machine Would Not Want to Commit Suicide

An argument put forward during the question session is that a human, when subjected to too much stress and difficult situations, can finally end his life, which a machine would not do.

As for emotions and lies, if the underlying argument is the conscious experience, that is to say that the suicidal act is an example of a strong conscious state, thenwe come back once again to the Turing test. If the argument is purely behavioral, one could perfectly imagine a machine which, given its objectives, prefers to self-destruct in certain situations.

B.2 A Machine Cannot Be Altruistic

A machine can perfectly demonstrate moral and altruistic principles, again according to its objectives. Simulations with robots that share resources with third parties show that such behaviors appear, and follow the predictions of evolutionary biology [14].

B.3 A Machine Is Replaceable

Finally, a last argument is that a machine can be replaced. Unlike a human being, one can substitute another copy of the same model if it is destroyed.

A first argument would be that the manufacturing process or the programming of a machine could ensure its uniqueness using random parameters that would be specific to it. But more fundamentally, even identical machines, if they are capable of learning, become unique after having existed long enough. They are the product of their interactions with the environment, which is specific to each one.

Like a pet, a robot that would live daily with a human, and with whom it would have a common past, would not be replaceable by a new copy.

References

[1] David J Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

[2] Ronan Collobert, Jason Weston, L'eon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[3] Daniel C Dennett. *Consciousness Explained*. Little, Brown and Co., 1991.

[4] Daniel C Dennett. The unimagined preposterousness of zombies. *Journal of Consciousness Studies*, 2(4):322–325, 1995.

[5] Junshui Ma, Robert P Sheridan, Andy Liaw, George E Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 55(2):263–274, 2015.

[6] Marvin Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon and Schuster, 2007.

[7] Sara Mitri, Dario Floreano, and Laurent Keller. The evolution of information suppression in communicating robots with conflicting interests. *Proceedings of the National Academy of Sciences*, 106(37):15786–15790, 2009.

[8] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974.

[9] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[10] John R Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980.

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 2014.

[12] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[13] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2013.

[14] Markus Waibel, Dario Floreano, and Laurent Keller. A quantitative test of hamilton's rule for the evolution of altruism. *PLoS Biology*, 9(5):e1000615, 2011.

[15] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 818–833. Springer, 2014.